AD —

UNCLASSIFIED

17605

U. S. Naval School of Aviation Medicine

U. S. NAVAL AIR STATION

PENSACOLA, FLORIDA

RESEARCH REPORT

MIDDLE CATEGORY ("?") RESPONSE: RELIABILITY AND RELATIONSHIP TO PERSONALITY AND INTELLIGENCE VARIABLES

REPORT NO. NM 001 077.01.01

U. S. Naval School of Aviation Medicine 16 June 1953

Project Number NM 001 077.01.01

Middle Category ("?") Response: Reliability and Relationship to Personality and Intelligence Variables.

Nathan Rosenberg and Carroll E. Izard, Tulane University, and Edwin P. Hollander, U. S. Naval School of Aviation Medicine, Pensacola, Florida.

14 pp. 5 tables UNCLASSIFIED

This study investigated the "?" response of objective personality tests with respect to its reliability and relationship to personality and intelligence. Four "?" scores were obtained for each of several hundred Naval Aviation Cadets from the three Guilford-Martin and the Bernreuter personality inventories. Results indicate that the tendency to respond by "?" is a reliable trait as evidenced by the relatively high intercorrelations between the number of "?'s" from the several personality tests used. Question mark scores do not correlate with scholastic aptitude (ACE) and educational level attained. They correlate significantly with many Guilford-Martin personality trait scores, but these correlations are artifacts of scoring procedures used.

1. Personality 2. Measurement

I. Rosenberg, Nathan II. Izard, Carroll E. III. Hollander, Edwin P.

U. S. Naval School of Aviation Medicine 16 June 1953

Project Number NM 001 077.01.01

Middle Category ("?") Response: Reliability and Relationship to Personality and Intelligence Variables.

Nathan Rosenberg and Carroll E. Izard, Tulane University, and Edwin P. Hollander, U. S. Naval School of Aviation Medicine, Pensacola, Florida.

14 pp. 5 tables UNCLASSIFIED

This study investigated the "?" response of objective personality tests with respect to its reliability and relationship to personality and intelligence. Four "?" scores were obtained for each of several hundred Naval Aviation Cadets from the three Guilford-Martin and the Bernreuter personality inventories. Results indicate that the tendency to respond by "?" is a reliable trait as evidenced by the relatively high intercorrelations between the number of "?'s" from the several personality tests used. Question mark scores do not correlate with scholastic aptitude (ACE) and educational level attained. They correlate significantly with many Guilford-Martin personality trait scores, but these correlations are artifacts of scoring procedures used.

1. Personality 2. Measurement

I. Rosenberg, Nathan II. Izard, Carroll E. III. Hollander, Edwin P.

U. S. Naval School of Aviation Medicine 16 June 1953

Project Number NM 001 077.01.01

Middle Category ("?") Response: Reliability and Relationship to Personality and Intelligence Variables.

Nathan Rosenberg and Carroll E. Izard, Tulane University, and Edwin P. Hollander, U. S. Naval School of Aviation Medicine, Pensacola, Florida.

14 pp. 5 tables UNCLASSIFIED

This study investigated the "?" response of objective personality tests with respect to its reliability and relationship to personality and intelligence. Four "?" scores were obtained for each of several hundred Naval Aviation Cadets from the three Guilford-Martin and the Bernreuter personality inventories. Results indicate that the tendency to respond by "?" is a reliable trait as evidenced by the relatively high intercorrelations between the number of "?'s" from the several personality tests used. Question mark scores do not correlate with scholastic aptitude (ACE) and educational level attained. They correlate significantly with many Guilford-Martin personality trait scores, but these correlations are artifacts of scoring procedures used.

1. Personality 2. Measurement

I. Rosenberg, Nathan II. Izard, Carroll E. III. Hollander, Edwin P.

U. S. Naval School of Aviation Medicine 16 June 1953

Project Number NM 001 077.01.01

Middle Category ("?") Response: Reliability and Relationship to Personality and Intelligence Variables.

Nathan Rosenberg and Carroll E. Izard, Tulane University, and Edwin P. Hollander, U. S. Naval School of Aviation Medicine, Pensacola, Florida.

14 pp. 5 tables UNCLASSIFIED

This study investigated the "?" response of objective personality tests with respect to its reliability and relationship to personality and intelligence. Four "?" scores were obtained for each of several hundred Naval Aviation Cadets from the three Guilford-Martin and the Bernreuter personality inventories. Results indicate that the tendency to respond by "?" is a reliable trait as evidenced by the relatively high intercorrelations between the number of "?'s" from the several personality tests used. Question mark scores do not correlate with scholastic aptitude (ACE) and educational level attained. They correlate significantly with many Guilford-Martin personality trait scores, but these correlations are artifacts of scoring procedures used.

1. Personality 2. Measurement

I. Rosenberg, Nathan II. Izard, Carroll E. III. Hollander, Edwin P.

U. S. NAVAL SCHOOL OF AVIATION MEDICINE
NAVAL AIR STATION
PENSACOLA, FLORIDA

JOINT PROJECT REPORT NUMBER 1

The Tulane University of Louisiana
Under ONR Project NR 154-098

and

U. S. Naval School of Aviation Medicine
Research Project Number NM 001 077.01.01

MIDDLE CATEGORY ("?") RESPONSE: RELIABILITY AND RELATIONSHIP TO PERSONALITY AND INTELLIGENCE VARIABLES

Report by

Nathan Rosenberg, M.A.
Carroll E. Izard, Ph.D.
and
Lieutenant (jg) Edwin P. Hollander, MSC, USNR

Approved by

Professor Cecil W. Mann
The Tulane University
and
Captain Ashton Graybiel, MC, USN
Director of Research
U. S. Naval School of Aviation Medicine

Released by

Captain James L. Holland, MC, USN
Commanding Officer

16 June 1953

SUMMARY

PROBLEM: In responding to items of psychological tests, subjects' answers are influenced by the form in which responses are presented. For example, personality test items customarily present a statement of behavior or feeling followed by response alternatives of "yes," "?," or "no." The tendency to answer "?" to personality items may be indicative of a personality trait itself, aside from the traits which the test was designed to measure. This study investigates the "?" response with respect to its reliability, and relationship to personality and intelligence.

SUBJECTS AND PROCEDURE: Three Guilford-Martin inventories were administered to 344 Naval Aviation Cadets. The Guilford-Martin tests purportedly measure 13 personality traits: GAMIN, STDCR, and OAgCo respectively. Bernreuter tests were also available for 277 of these subjects. Thirteen Guilford-Martin trait scores, ACE test scores, and years of schooling completed were used as independent variables. Four "?" scores were obtained by summing the number of "?" responses on each personality test. Intercorrelations were computed for these four scores, and correlations between selected combinations of scores. From this analysis, Bernreuter "?" scores and the sum of "?" scores from three Guilford-Martin tests were defined as dependent variables. These two scores were then correlated with Guilford-Martin trait scores, ACE scores, and years of schooling completed.

RESULTS: (1) The mean number of "?'s" was equal to about eight percent of the number of personality items.

(2) The distribution of "?" scores approximated a J-curve, one-half of the normal distribution.

(3) The reliability for Bernreuter "?" scores was estimated as .64; for a single Guilford-Martin test as .72; for the sum from three Guilford-Martin inventories in the low eighties.

(4) Question mark scores were independent of ACE and educational level attained within the restricted range of scholastic aptitude studied.

(5) Bernreuter and Guilford-Martin "?" scores correlated significantly, and about equally, with ten Guilford-Martin traits: GAMIN, OAgCo, and TR of test STDCR. These correlations were judged spurious, however, because of the method of scoring "?'s" utilized in Guilford-Martin tests.

CONCLUSIONS: Subjects who frequently use the "?" response to personality items of one test tend to use it often on other similar tests. Thus, the number of "?" responses is a measure of a reliable trait. The subject's intelligence does not appear to influence the number of "?" responses he will use. The tendency to use the "?" category was found to be correlated with many specific personality traits. These correlations were interpreted as an artifact of the method of scoring personality tests,

rather than representing the psychological correlates of people who respond by "?." Independent measures of personality are needed to find such correlates of "?" responses.

I. INTRODUCTION

A. History of the Problem

Many investigators have been concerned directly or indirectly with response sets. Cronbach (4) defines response set as "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in a different form." For example, observations of variability in such factors as speed and accuracy on objective tests and productivity on essay tests are commonplace. This report presents preliminary findings concerning a long discussed problem for which little empirical evidence has been available.

In its broader dimensions, response sets have been discussed in relation to topics varying from constant errors in psychophysics (4) to work simplification by means of time and motion studies (14). The use of the middle category, one type of response set, dates back to Titchener who was concerned with equality judgments in psychophysical research (18). A typical experiment in psychophysical research consists of having subjects judge whether one weight is heavier, lighter, or equal in weight to another. Titchener discussed the issue of retaining the middle category (e.g. equal in weight) versus its elimination, without clearly resolving it. In 1907, Angell (2) noted that subjects exhibited differences in their use of equality judgments, and he felt "that this difference corresponds to the difference between deliberate and impulsive temperaments." Fernberger (6) found that different instructions produce significantly different numbers of equality judgments in psychophysical experiments. Based on this finding and the results of previous research, he concluded that equality judgments are dependent upon form of instructions, subjects' attitudes, and their basic temperament. He felt it desirable to retain the psychophysical method which utilized a middle category. Woodworth (19) reviewed the debate about middle category retention versus elimination and concluded that from the laboratory point of view there was no real basis for favoring one method over the other.

B. Recent Research Relating to the Problem

In selecting one of the response categories of objective-type questionnaires, a subject is confronted with what is analagous to a psychophysical judgment. In addition to this, there also exists a semantic problem, as shown in a study by Mosier (9). He concluded that there were reliable differences in subjects' interpretation of words commonly used in interest, attitude, and personality tests. The meanings assigned by students to such words as "frequently," "indifferent," and "desirable" differed significantly. He found that students preferred "good" to "better," and "bad" to "worse," as shown by their ratings of these words.

Some of the research in this area relates to the general problem of the number and nature of response categories to be used. Osgood (10) showed that on a seven point scale some subjects predominantly used 1 and 7, some used 1, 4, and 7, while some used the whole scale. Remmers and Sageser (11) showed that within limits the more choices on a multiple-choice attitude scale the greater the test reliability.

In an article concerned with the effect of response sets on reliability and validity, Cronbach (4) listed the following response sets: (1) Tendency to gamble; (2) definition of judgment categories; (3) inclusiveness of response; (4) bias, acquiescence, tendency to agree; (5) speed versus accuracy. He presented evidence from the literature that each of these types of response sets was a reliable trait in certain test situations. He felt that the existence of still other response sets was likely.

Most of the studies in this area have aimed at establishing one or another of the response sets as a stable trait or factor which could be measured reliably. In one of the first and more extensive studies along this line, Lorge (8) found positive correlations between corresponding answer categories of the Bernreuter, the Strong Vocational Interest Blank, and a Thorndike and a Thurstone attitude scale. For example, he found a positive correlation between the number of "?" responses on the Bernreuter, the number of "?" responses on the Thurstone, the number of "I" responses on the Strong, and the number of "5" responses on the Thorndike. The correlation coefficients were not reported.

Lorge inferred from his findings (1) that the method of rating items introduced a special effect which he considered a halo effect; (2) "that the tendency to respond by 'yes's,' 'no's,' '?'s,' or similar rubrics may be symptomatic of a special aspect of personality." The first inference implies that the desired approach to the problem is elimination or control of response set through improved test construction. This conclusion has been reached in most of the subsequent studies. Lorge's second inference implies that measures of response sets may represent measures of personality. There have been few empirical studies relating to this inference.

Argument for trying to eliminate or control response-set variability is engendered by evidence demonstrating the effects of this source of variation on test reliability and test validity. Lentz (7) reported that acquiescence, or tendency to agree, was a potent factor in lowering reliability of personality measurements and pointed out the need for controlling this factor. Cronbach (3) investigated the factor of acquiescence and its effect on the reliability and validity of a series of true-false tests. The acquiescence factor had test-test reliability coefficients which ranged from .36 to .61, all of which were significant at the .01 level of confidence. The reliability coefficients of "false" scores (based on items marked false) were generally greater than those for "true" scores (based on items marked true). In three cases out of a total of ten, C. R.'s between corresponding reliabilities for "true" and "false" scores were significant.

Using course grades derived from tests other than true-false as a criterion, the following validity coefficients were reported for the "true," "false," and total scores of two true-false tests:

	"True"	"False"	Total
Test 1	.222	.666	.670
Test 2	.319	.700	.598

Cronbach's study presents cogent evidence for the existence of the acquiescence factor and its effects on the reliability and validity of true-false tests.

Rundquist (13) showed that form of statement, a correlative of response set, affected the validity of items on a personality test of the personal inventory type. Scores on "acceptable" (positively state items) were less valid than scores on "unacceptable" (negatively) stated items.

After reviewing numerous studies demonstrating various response sets, Cronbach (5) stated that response sets generally tend "to reduce the saturation of a test and to limit its possible validity." He recommended that response sets be avoided " with the occasional exception of some tests measuring carefulness or other personality traits which are psychologically similar to response sets." Whether a response set and a personality trait are "psychologically similar" is a matter of hypothesis, however. Empirical evidence for the relationship between response set and personality is rare. Two studies which move in this direction are those of Swineford (16) and Lorge (8).

Swineford (16) found evidence for the existence of "tendency to gamble" as a stable factor. This gambling tendency is measured by allowing subjects to assign 1, 2, or 4 points to objective test items which have right and wrong answers. This technique is an indirect method of assessing the certainty which a subject ascribes to his answers. The intercorrelation of G (gambling) scores from four different tests ranged from .201 to .798, with a multiple R of .847. The distributions of the G scores were positively skewed; none of them approached normality. The only evidence for G as a "personality trait" was that it was so named, and was not correlated with ability factors.

Smith and Tyler (15) found that the tendency to use intermediate rather than more extreme scale positions could be used as a reliable index of students' behavior with respect to their "caution" in drawing conclusions. They reported a test-test reliability coefficient of .85 for the caution factor.

Rundquist (13) considered the possible meaningfulness of the response set termed "the tendency to take extreme scale positions." He tested 111 factory girls with separate series of personality and interest items. He found that the tendency to take extreme scale positions correlated .40 between interest and personality items. In view of this relative-

ly low reliability, he felt that the response set reflected situational factors rather than anything basic to personality. With regard to the type of personality and interest items used, he regarded the elimination of response set as more profitable than attempting to measure it. The fact that Rundquist correlated the total for both extremes of the scale on one test with the total for both extremes on the other may have suppressed reliability. There is reason to believe that a separate response set operates for each extreme of a given scale.

The research reviewed in this section indicated the existence of various types of response sets. However, the question of what to do about a response set once it has been found remains unanswered. Some authors (3,7,12) have shown that in certain test situations response sets adversely affect reliability and validity. Other investigators (15,16) working with different tests, have indicated that response sets are valuable indices of certain aspects of personality. Apparently, the dissimilarity of these findings stem more from the differences in interests and purposes of the investigators than anything else. It appears that the question of what to do with a response set may have to be answered separately for each psychological test. Whenever an existing test is shown to yield a stable measure of a response set, the problem of dealing with it must be resolved by weighing the effects on reliability and validity on the one hand and the intrinsic value of the response set measure as an index of personality on the other.

II. STATEMENT OF PRESENT PROBLEM

This investigation is concerned primarily with a study of the relationships of the middle-category ("?") response set to certain personality variables. At the same time this research throws light on the question of the existence and stability of this particular response set. In particular, the following hypotheses will be tested:

(1) That there is a response set which predisposes some individuals to give a greater number of question mark responses than do others.

(2) That there are significant correlations between the relative number of question mark responses which individuals give on one test and the number which they give on other tests.

(3) That a disposition toward giving question mark responses may be shown to be related to certain dimensions of personality.

III. METHOD

Guilford-Martin and Bernreuter personality inventory scores for several hundred Naval Aviation Cadets at Pensacola, Florida were available for study by virtue of their use in another research project.* The Guilford-Martin inventories used, with the definition of personality trait scores included:

1. An Inventory of Factors STDCR (Test STDCR)

   a. S ___ Social introversion - extraversion
   b. T ___ Thinking introversion - extraversion
   c. D ___ Depression
   d. C ___ Cycloid disposition
   e. R ___ Rhathymia

2. The Guilford-Martin Inventory of Factors GAMIN (Test GAMIN)

   a. G ___ General pressure for overt activity
   b. A ___ Ascendancy
   c. M ___ Masculinity
   d. I ___ Lack of inferiority feelings
   e. N ___ Lack of nervous tenseness

3. Guilford-Martin Personnel Inventory (Test OAgCo)

   a. O ___ Objectivity
   b. Ag___ Agreeableness
   c. Co___ Cooperativeness

The papers for these cadets were scored by keying the "?" response with a weight of one. Three "?" scores are thus obtained for the 344 cadets on the three tests of the Guilford-Martin series and one score for 277 of these subjects on the Bernreuter. These scores represent the number of "?" responses given by a subject on each of the four personality tests administered. These data along with ACE Quantitative, Linguistic, and Total Scores, educational level, and the thirteen personality trait scores derived from the Guilford-Martin inventories were punched into IBM cards.

IV. RESULTS

1. The mean number of "?" responses is equal to about eight percent of the number of personality items. Variability of "?" scores is considerable. (See Table I).

2. The distribution of "?" scores approximates a J-curve. (See Table II).

*The authors are grateful to Drs. Richard Trumbull and John Manhold for making these data available.

3. The tendency to use the "?" category is a reliable trait as shown by the intercorrelations for "?" scores from various personality tests. (See Table III).

4. Question mark scores do not correlate significantly with scholastic aptitude and educational level attained within the restricted range of scholastic aptitude studied. Bernreuter and Guilford-Martin "?" scores correlate significantly, and about equally, with ten Guilford-Martin personality traits: CAMIN, OAgCo, and TR of Test STDCR. (See Table IV).

V. DISCUSSION OF RESULTS

A. Reliability and Related Statistics

The variability and the nature of the distribution are among the first problems to consider in describing a trait. Thus, the extent to which the "?" category is used, and the nature of the distribution of "?" scores were presented first in Tables I and II. From these data, it is noted that the mean number of "?" responses for this group was about equal to 8% of the number of items of the Guilford-Martin and Bernreuter tests. Furthermore, subjects differ considerably in the extent to which this category is used, as shown by the standard deviations and frequency distribution presented. This distribution is positively skewed; the shape approximating closely the J-curve, half of a normal distribution.

Having established that the variability in "?" scores is considerable, the problem of reliability arises. With "?" scores from four separate tests to be correlated with many other variables, it becomes a practical as well as theoretical issue to consider the best single composite score that can be derived from the many possible combinations of scores. The correlations reported in Table III were calculated with this in mind. The blank cells in the matrix involve correlations of a single test score with a sum which would be based in part on itself. This would result in reliabilities spuriously high, and make their interpretation difficult.

Certain relationships are immediately apparent from this table. The intercorrelations between the three Guilford-Martin tests are slightly higher than between the Bernreuter and Guilford-Martin tests. This discrepancy could arise from many factors, but one of the first to consider is the difference between the number of items in the Bernreuter and the Guilford-Martin inventories. The former consists of 125 items and the latter of 511. We can approximate the number of items on each of the Guilford-Martin tests as 175. Applying the Spearman-Brown prophecy formula to the 125 item Bernreuter test, whose reliability is estimated as .64, the increased length to 175 items yields a calculated reliability of .71. This value approximates closely the observed reliability of .72 estimated for the Guilford-Martin tests.

Many factors could account for the observed correlations, but other things being equal, the most important factor influencing reliability

is the sampling of items. From the above results, Bernreuter "?" scores would seem to have as much in common with Guilford-Martin "?" scores as the latter have with each other. This suggests that differences in item content between the Bernreuter and Guilford-Martin are equally unimportant in the production of "?" responses.

The reliability for the total "?" score obtained from the three Guilford-Martin inventories was desired since these tests are commonly administered together. The reliability for a single Guilford-Martin test can be estimated from the intercorrelations between the three inventories. The Spearman-Brown formula for tripling the length of a test is then applied to this reliability estimate for a single inventory. This value calculated from the Spearman-Brown is a maximum estimate for the reliability of the total "?" score from the Guilford-Martin inventories.

The reliability for a single Guilford-Martin test is estimated as .72, the median value for the intercorrelations (.70, .72, and .76) of the three Guilford-Martin tests. The calculated value from the Spearman-Brown formula is .89, an estimate of the maximum reliability when a test has been tripled in length by the addition of comparable items.

A minimum estimate of the reliability for total "?" scores is obtained by considering the correlation between "?" scores for a single Guilford-Martin test and the sum of the remaining two inventories. This reliability estimate is less than that for tripling the length of a test. The estimates found in Table III are .76, .78, and .80; the median value of .78 is considered as the best minimum reliability estimate.

From the maximum and minimum estimates obtained, it is concluded that the reliability for total "?" scores from the Guilford-Martin is in the low eighties.

B. Correlations With Other Variables

The correlations reported in Table III indicate that a stable trait is being measured for the samples of behavior observed. These results are in keeping with previous research in this area. An important hypothesis remains to be tested. That is, are there significant relationships between "?" scores and other variables, particularly those in the personality realm? It was postulated that the tendency to use the "?" category is a response set indicative of personality trends.

There remains to be demonstrated that significant correlations exist between "?" scores and personality variables. The only data available at this writing in the personality area include scores on personality traits from the Guilford-Martin itself. One other hypothesis for which data were available concerns the relationship between scholastic aptitude, educational level, and "?" scores. Table IV lists the correlations for these variables.

Because personality trait scores from the Guilford-Martin may be contaminated by the number of "?" responses, separate correlations were computed for Bernreuter "?" scores. The sum of all "?" responses for the four tests were also correlated with Guilford-Martin trait variables to note any increase in correlation due to increased reliability by virtue of increased length. The issue of contamination of Guilford-Martin personality trait scores by the "?" category will be discussed at greater length shortly.

It will be noted in Table IV that scholastic aptitude (as measured by the ACE) and educational level (number of years schooling completed) do not correlate significantly with the "?" variable. This result is consistent with the findings of Swineford (16), referred to earlier. The independence of "?" scores from variables of education and scholastic aptitude is a very useful property for predictive problems involving multiple correlation. This implies that significant correlations may be found between "?" scores and other predictors of the criterion investigated.

In the personality realm, significant correlations are noted between "?" scores and all Guilford-Martin trait scores from Test GAMIN and OAgCo. In addition, traits T and R from test STDCR correlate significantly with the "?" variables. No great change in the magnitude of these correlations is noted when either the total Guilford-Martin "?" scores are used, or the Bernreuter, or the sum of both. There is some tendency for the total sum to correlate higher, but the differences noted would not seem to warrant the additional testing time required.

On the basis of these correlations, it might be concluded that individuals who tend to respond with many "?" responses tend to be less objective (O), agreeable (Ag), cooperative (Co), active (G), ascendant (A), masculine (M), self-confident (I), free from neurotic tendencies (N), reflective (T), and impulsive (R).

It is of interest to note that the three different "?" scores used to correlate with Guilford-Martin trait scores yield comparable results. There is generally some decrease in the size of these correlations for the Bernreuter. However, the same traits are revealed as significant. This is of practical importance since testing time available could become crucial in other studies involving the "?" variable. It would appear that for the purpose of demonstrating the existence of correlation between "?" scores and other variables, as well as a general indication of the strength of the relationship, a single personality inventory is almost as useful as three.

The finding that "?" scores correlate negatively with Guilford-Martin trait scores is consistent with the expected personality pattern of subjects tending to respond with many "?'s." Nevertheless, these results cannot be taken at face value. All Guilford-Martin trait scores are influenced by how many "?" responses are selected by a subject. If all items of the Guilford-Martin are marked "?," the results indicated in Table V are obtained.

In the case of Tests GAMIN and OAgCo, subjects must respond either "yes" or "no" to receive points toward the traits being measured. This is also true for traits T and R of Test STDCR. The more "?" responses a subject checks the more likely he will receive a low raw score on all of the above traits. In the case of traits S, D, and C of Test STDCR, a subject may accumulate points toward these traits by marking the "?" category. For these traits two subjects may have identical raw scores, one by virtue of many "?" responses, the other by "yes" and "no" responses which counteract each other. Thus, the correlations actually observed are explicable in view of these considerations. It will be recalled that negative correlations were obtained for trait scores where the "?" response was not scored, and chance correlations for traits (S, D, and C) where the "?" response was an important factor in the trait scores. It is these facts that make difficult the interpretation of the significant correlations obtained between personality traits and "?" responses.

Bernreuter "?" scores and Guilford-Martin trait scores are independent measures in the sense that Bernreuter "?" scores do not enter into the score for Guilford-Martin traits. It appeared at the outset of this study that the correlations for the Bernreuter would be of crucial importance. However, the Bernreuter correlates .72 with the total Guilford-Martin "?" score. The correlation of .72 suggests that Bernreuter "?" scores would be expected to correlate with the same variables as do Guilford-Martin "?" scores.

C. THEORETICAL IMPLICATIONS

One channel of thinking about the "?" response emerges from the J-curve found for the distribution of scores. In social psychology curves of this type have been described by Allport (1) as conformity curves. Social situations which demand conformity yield distributions of scores when measured which are highly skewed, the modal behavior approaching the cultural norm. The classic example cited is the behavior of motorists at an intersection who are confronted by either red lights, stop signs, or a traffic officer. Most motorists in such situations will stop completely, some will go very slow or slightly slow, and a few will not reduce their speed at all.

By analogy, behavior in answering personality items could be a reflection of cultural conformity. The subject who responds with very many "?" responses may be going through the stop sign, so to speak. In effect, he may be avoiding the test through the refuge of the "?" response. From this analysis, each personality item may be a reflection of a cultural norm. The major difference between behavior in the stop sign situation and toward personality items is that behavior in the first case is defined explicitly, in the second implicitly.

We could define operationally the cultural norm for any item as the proportion of people responding in the majority direction of either "yes"

or "no." Where the majority for "yes" or "no" responses is very decisive, these segments of behavior represent cultural norms clearly crystallized for that group. The psychological pattern inferred from the content of such items might be interpreted as the cultural definition of adjustment. If such a framework can be formed, a large "?" score, ipso facto, represent non-adjustment according to cultural conformity.

Although the suggested approach outlined above may be worthwhile, the contamination of personality trait scores with the "?" response makes necessary the adoption of completely independent criteria to which "?" scores may be correlated. Preliminary analysis by the writers show, for example, that "?" scores are not related to "buddy ratings," scores obtained by peer nominations for leadership qualities. Further work along this line will be reported later. What is needed are many such independent measures which seem to be based in great part upon personality factors.

Guilford himself must recognize the problem posed by the "?" category. In the Guilford-Zimmerman personality test, a revision of the Guilford-Martin, the directions now stress that subjects should avoid using "?" responses unless absolutely necessary.

It may turn out that the "?" response in itself is not a significant variable. However, more fruitful results may be obtained from standard personality tests such as the Guilford-Martin when trait scores are corrected for the number of "?" responses. Thurstone (17) has already recommended just such a procedure in his manual for the Thurstone Temperament Schedule. He calls for the two new types of scores: (1) the number of "?" responses made on the Temperament Schedule and (2) a score for each personality area which is twice the number of correct responses plus the number of "?" responses. Thurstone calls these "Experimental Uncertainty Scores," and postulates that the first of these may indicate lack of self-confidence, or insecurity in self appraisal. The second score makes an adjustment in trait scores so as to differentiate between subjects who otherwise might receive the same raw score.

It would seem worthwhile when using objective personality tests as predictors to develop three scores. The first would be the usual trait scores derived from the test. The second would be the number of "?" responses. The third would be the trait scores corrected for the number of "?" responses in each trait. The adjustment made could be the one recommended by Thurstone, or a similar correction.

There remains one further issue to explore. In considering the three responses to personality items, there are several assumptions which can be made with respect to the underlying continuum. The usual one made is that the "?" response lies between "yes" and "no" on a continuum of judgment. For the purpose of exposition, let us consider one item chosen arbitrarily from the Guilford-Martin inventories: "Do you like to speak in public?" Under the above assumption, the continuum implied is the judgment by a subject with regard to his liking-to-speak-in-public. If his judgment is that

this behavioral gestalt is typical of him, he responds "yes." If it is not, he responds "no." By answering with the "?" response, he implies that his judgment is uncertain. The crucial feature of this kind of continuum is that subjects should perceive the item content similarly.

Another possible assumption is that "yes" and "no" represent opposite poles of a continuum, while the "?" is not on this continuum at all. Stated another way, the "?" category is qualitatively different from the other two. This assumption implies that subjects who respond "?" to an item are not all reacting alike to the gestalt-like feeling associated with the item content. Some of them may respond "?" because the item does not apply to them, they do not understand its meaning, or they agree to some part (like to speak) but disagree with another part (in public).

Similar reasoning may hold with respect to "yes" and "no" responses; that is, they too may represent qualitatively different responses. Other research has presented evidence for the existence of response sets for these categories. Since these types of response sets have not been explored in this paper, crucial data are lacking for the point in question.

It is not feasible to record the thought processes of subjects as they respond to personality items. Yet, the distribution of scores for each of the response categories may make possible inferences about the assumptions in question. It has been reported here that the distribution of scores for the "?" response is highly skewed, approximating the J-curve. Further research comparing the reliability and distribution of scores for "yes" and "no" response sets to those for "?" may shed light on this issue.

In the first portion of the theoretical implications, we have discussed a method of exploring the meaning of item responses in their present form. This last section suggests that the results of considering the different assumptions underlying the response continua may force a revision in the type of responses available to the subjects. One such revision possible is the abandoning of the middle category. An alternative is to explore the meaning of "?" scores when this category is retained.

Further work is being done on the relationship of middle category scores to independent measures of personality. These results will be reported later.

VI. CONCLUSIONS

With respect to the hypotheses formulated in the statement of the problem, it may be concluded from the results presented here that:

1. There is a response set which predisposes some individuals to respond with a greater number of "?" responses than do others. Furthermore, the distribution of such scores approximates closely a J-curve.

2. The reliability of "?" scores as shown by the correlations between several tests indicates a stable trait is being measured.

3. Question mark scores are:

   a. Independent of scholastic aptitude (ACE) and educational level attained.

   b. Related to Guilford-Martin personality trait scores GAMIN, OAgCo, and TR of Test STDCR. The evidence from this study indicates that the significant correlations found are spurious by virtue of the procedures followed in scoring "?" responses on the Guilford-Martin inventories.

4. It is also concluded that personality trait scores derived from objective personality tests should be adjusted by trait for the number of "?" responses. The importance of this adjustment will be a function of the number of times the "?" category is used and the scoring procedures followed with respect to this category.

BIBLIOGRAPHY

1. Allport, F. H. The J-curve hypothesis of conforming behavior. J. soc. Psychol., 1934, 5, 141-183.

2. Angell, F. On the judgment of 'like.' Amer. J. Psychol., 1907, 18, 253-260.

3. Cronbach, L. J. Studies of acquiescence as a factor in the true-false test. J. educ. Psychol., 1942, 33, 401-415.

4. Cronbach, L. J. Response sets and test validity. Educ. Psychol. Measmt, 1946, 6, 475-494.

5. Cronbach, L. J. Further evidence on response sets and test design. Educ. Psychol. Measmt, 1950, 10, 3-31.

6. Fernberger, S. W. The use of equality judgments in psychophysical procedures. Psychol. Rev., 1930, 37, 106-112.

7. Lentz, T. F. Acquiescence as a factor in the measurement of personality. Psychol. Bull., 1938, 35, 659.

8. Lorge, I. Gen-like: Halo or reality. Psychol. Bull., 1937, 34, 545-546.

9. Mosier, C. I. A psychometric study of meaning. J. soc. Psychol., 1941, 13, 123-140.

10. Osgood, C. E. Ease of individual judgment-process in relation to polarization of attitudes in culture. J. soc. Psychol., 1941, 14, 403-418.

11. Remmers, H. H. and Sageser, H. W. Reliability of a multiple-choice measuring instrument as a function of the Spearman-Brown formula. J. educ. Psychol., 1941, 32, 445-451.

12. Rundquist, E. A. Form of statement in personality measurement. J. educ. Psychol., 1940, 31, 135-147.

13. Rundquist, E. A. Response sets: A note on consistency in taking extreme positions. Educ. Psychol. Measmt, 1950, 10, 97-99.

14. Seashore, R. H. Work methods: An often neglected factor underlying individual differences. Psychol. Rev., 1939, 46, 123-141.

15. Smith, E. R. and Tyler, R. W. Appraising and recording student progress. New York: Harper and Brothers, 1942, III.

16. Swineford, F. Analysis of a personality trait. J. educ. Psychol., 1941, 32, 438-444.

17. Thurstone, L. L. Examiner manual for the Thurstone Temperament Schedule. Chicago: Science Research Associates, 1950.

18. Titchener, E. B. Experimental psychology. New York: MacMillan Co., 1905, II.

19. Woodworth, R. S. Experimental psychology. New York: Holt and Co., 1938.

TABLE I

MEANS, STANDARD DEVIATIONS, AND
PERCENTAGES OF QUESTION MARK SCORES ON FOUR PERSONALITY TESTS

	Test	No. of Items	No. of Subjects	Mean No. of "?"	Standard Deviation	% of No. of Items
1.	Guilford-Martin (GAMIN)	186	344	13.30	13.18	7.15
2.	Guilford-Martin (STDCR)	175	344	13.21	13.84	7.55
3.	Guilford-Martin (OAgCo)	150	344	11.60	11.33	7.73
4.	Bernreuter	125	277	12.33	11.46	9.86
5.	Guilford-Martin Sum (1, 2, and 3 above)	511	337*	36.38	35.26	7.12
6.	Bernreuter and Guilford-Martin Sum (4 and 5 above)	636	273*	49.61	44.93	7.80

*For IBM analyses, subjects were eliminated with incomplete data on the remaining tests.

TABLE II

FREQUENCY DISTRIBUTION FOR SUM OF "?'s" ON THREE GUILFORD-MARTIN TESTS (N = 337)

Question Mark Score	Frequency
0-9	102
10-19	41
20-29	39
30-39	35
40-49	29
50-59	27
60-69	18
70-79	7
80-89	10
90-99	10
100-109	5
110-119	2
120-129	1
130-139	5
140-149	3
150-159	0
160-169	0
170-179	1
180-189	0
190-199	0
200-209	2

TABLE III

INTERCORRELATIONS BETWEEN NUMBER OF QUESTION MARK RESPONSES FOR GUILFORD-MARTIN AND BERNREUTER PERSONALITY INVENTORIES*
(N = 277)

		1.	2.	3.	4.	5.	6.	7.	8.
1.	Bernreuter	---							
2.	OAgCo	.64	---						
3.	GAMIN	.64	.76	---					
4.	STDCR	.67	.70	.72	---				
5.	OAgCo and GAMIN	.68	---	---	.76	---			
6.	OAgCo and STDCR	.72	---	.80	---	---	---		
7.	GAMIN and STDCR	.70	.78	---	---	---	---	---	
8.	Total Guilford-Martin Sum	.72	.89	.91	.91	---	---	---	---

*Blanks represent correlations of a single score with a sum based in part on that score. Such correlations are spuriously high.

TABLE IV

CORRELATIONS BETWEEN QUESTION MARK SCORES AND GUILFORD-MARTIN PERSONALITY TRAIT SCORES, ACE, AND EDUCATION

		Total Guilford-Martin "?'s" (N = 337)	Bernreuter "?'s" (N = 273)	Sum of Guilford-Martin and Bernreuter "?'s" (N = 273)
1.	ACE-Q	.04	-.07	.00
2.	ACE-L	.03	.00	.03
3.	ACE-Total	.04	-.02	.03
4.	Education	.02	-.01	.01
5.	S	.04	.05	.09
6.	T	-.16	-.22	-.18
7.	D	-.01	-.05	.03
8.	C	-.06	-.10	-.03
9.	R	-.39	-.33	-.42
10.	O	-.30	-.17	-.30
11.	Co	-.26	-.17	-.24
12.	Ag	-.24	-.15	-.26
13.	G	-.29	-.26	-.30
14.	A	-.22	-.15	-.22
15.	M	-.36	-.23	-.35
16.	I	-.30	-.19	-.30
17.	N	-.22	-.12	-.23

TABLE V

HOW "?" RESPONSES ENTER INTO TRAIT SCORES

	Test	Total No. Items	No. Items Scored For "?" By Trait	C-Score Obtained
1.	STDCR	175	S-27; T-0; D-17; C-14; R-0	S-3; T-10; D-6; C-8; R-0
2.	OAgCo	150	O-3; Ag-5; Co-6	All C scores are 0
3.	GAMIN	186	None	All C scores are 0